冗長情報を利用した手書き住所読み取り知識処理の一評価

4H－9

下村　秀樹
NEC 情報メディア研究所

1. はじめに

一般に，記入枠制限のない文字列の読み取りでは単独文字の切り出しや認識が困難であり，読み取り対象依存の知識を用いた「知識処理」による文字認識候補の選択／修正が不可欠である．住所に関しては従来から，(1)地名の階層関係制約，(2)丁目以下が数字と区切り記号の並びから成るという規則，(3)地名ごとの丁目以下の値範囲制限，などの知識を使った「知識処理」が提案されている[1][2]．また，その一部は帳票読み取り，郵便宛名読み取りなどですでに実用化されている．しかし，個別文字の激しい変形や隣接文字の接触・入り組みが起こった場合，特にそれが丁目以下の数字部分に発生した場合には，文字候補を選択／修正するに十分な情報がなく，誤読あるいは棄却を正読とすることが非常に難しい．

これに対し文献[3]では，従来とは別角度からの知識処理のアプローチとして，読み取り対象文字列を構成する要素単語間の冗長情報を利用した方式を提案した．この方式は，文字認識が不完全であっても，そこからあいまい一致で抽出した要素単語群と正解テーブルとを照合することにより，認識結果を補正できるという頑健性を持つ．

本稿では，この提案方式の効果に関する一評価実験の結果を報告する．

2. 単語冗長情報を利用した知識処理

提案手法の概略を図１に示す．まず入力パタンに対し個別文字切り出し・認識を行い文字認識候補を得る．候補は切り出し・認識ともに複数出る．次にその候補文字群で単語辞書を検索し，読み取り対象を構成する要素単語候補を抽出する．単語検索では候補文字群と単語の完全一致ではなく，一部に文字の欠落があっても候補として挙げる「あいまい照合（虫食い照合）」

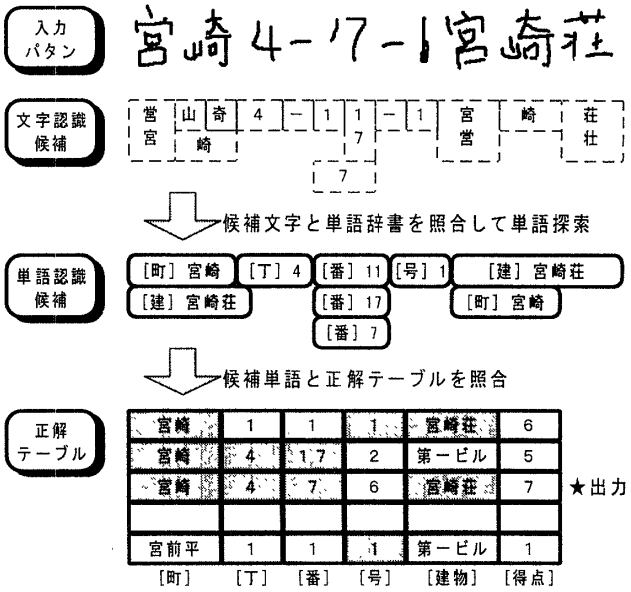

図1　単語間の冗長性を利用した知識処理

を行う．最後に要素単語候補群と正解を格納したテーブルを照合し，最もよく一致した正解テーブル中のレコードを読み取り結果として出力する．この方式の特長は，要素単語の認識誤りや棄却があっても，他の単語と正解テーブルによって補正できるという頑健性にある．例えば図１では，文字認識で丁目以下の「号」の部分を「１」と誤っているが，正解テーブルと一致した単語文字数（図１中の[得点]）が最大のレコードを読み取り結果とすることで，最終的に「６」に補正することができている．もちろん，得点の定義はこれ以外にも考えられる．

冗長情報による認識候補選択／補正の手法としては，記入枠ありで氏名漢字とふりがなを使った例などが報告されている[4]．これに対し本稿で述べている手法は，いったん単語を抽出した後に正解テーブルと照合することで，記入枠なしの場合にも計算量を抑えて処理を行うことを可能としている．また，この手法ではあいまい照合（不完全一致）を許して文字候補群から単語候補を探索することが必須となるが，これについては文献[5]で筆者らが提案したアルゴリズムが有効である．

An evaluation of a knowledge-based processing method using redundant information for handwritten address reading
Hideki SHIMOMURA
Information Technology Res. Labs., NEC corp.

3. 評価実験

東京都内の1つの区内の住所を対象とし，地名・丁目・番地・号・建物名を要素単語として，提案手法の評価実験を行った．

実験システムの構成を図2に示す．まず文献[2]の従来方式により地名と街区（丁目・番地・号）の制約条件に基づく住所読み取りを行い，次にその結果から要素単語を切り出して正解テーブルと照合することで提案手法を実現した．文献[2]では建物名を認識していないが，実験では文献[5]のアルゴリズムを流用して建物名の単語認識も別に行った．

評価画像は，住宅地図と電話帳を参照して手作業で作成した後，従来方式において町名候補が少なくとも1つは出力された250枚を選んだ．知識処理の効果を調べる実験なので，町名が全く読めないような質の悪い画像を評価から除外する意図でこの条件を付けた．正解テーブルは，評価画像に記載されている住所，建物名をそのまま格納したものに，ダミーを250件加えた500件で構成した．したがって，評価画像に対する正解は，必ず正解テーブルに格納されているという前提での実験となる．住所を構成する要素単語の種類は，建物名500，地名約50，それと丁目以下の数値である．

表1に，従来方式を単独で用いた場合と提案手法を用いた場合の読み取り性能比較を示す．判定は次の通りである．

「正読」：町名・丁目・番地・号をすべて正しく出力したもの

「部分正」：認識不良や競合が発生したため号までは一意に定まらなかったが出力した部分に誤りはないもの

「誤読」：出力のどこかに誤りがあるもの

「棄却」：競合により町名が一意に定まらず何も出力しなかったもの

表1から，従来方式では認識不良や単語候補競合のため46.0%であった正読率が，提案方式により92.8%に大きく改善されたことがわかる．冗長情報を利用した知識処理方式の住所読み取りでの有効性が確認されたといえる．ただし提案手法でも，誤読した要素単語の組み合わせが偶然正解テーブルの別レコードと一致した，あるいは同点の候補が競合したケースは，正読とすることができなかった．また，建物名を誤って読んだことに起因する「部分正」から「誤読」への改悪も1例あった．

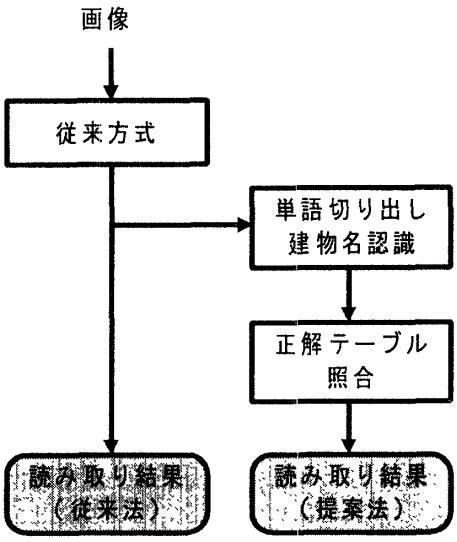

図2　実験システムの構成

4. おわりに

住所を構成する要素単語の冗長性を利用した知識処理方式の評価を行い，その有効性を示した．実験から，提案手法の基本的な妥当性は検証できたと考える．

しかし，正解テーブルのデータ量，認識対象単語数などによって，性能は大きく変動すると思われる．また，正解テーブルに登録もれや誤りがある場合についての対処なども興味深い問題である．今後はそれらについても検討していきたい．

参考文献

[1]清野他:自由記載住所文字列に対する知識処理,1989信学全大,D-465,1989

[2]下村他:効率的探索とトップダウン的検証を組み合わせた手書き住所読み取り知識処理,52回情処全大,4G-4,1996

[3]下村:要素単語の相互チェックに基づく手書き文字列認識知識処理,53回情処全大,2N-11,1996

[4]浅見他:ふりがなを利用して認識率を上げた手書き漢字OCR,日経エレクトロニクス,1998.10.17

[5]福島他:手書き文字列読み取りのための単語列探索アルゴリズム－文字タグ法－,情処論文,Vol37,No.6,1996

表1　従来法と提案法の性能比較（250画像）

	正読	部分正	誤読	棄却
従来法	115 (46.0%)	104 (41.6%)	28 (10.4%)	3 (1.2%)
提案法	232 (92.8%)	12 (4.8%)	6 (2.4%)	0 (0.0%)